Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification

Joel TETREAULT[1] *Daniel BLANCHARD*[1]
Aoife CAHILL[1] *Martin CHODOROW*[2]
(1) Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA
(2) Hunter College and the Graduate Center, City University of New York, New York, NY 10065, USA
tetreaul@gmail.com, dblanchard@ets.org, acahill@ets.org,
martin.chodorow@hunter.cuny.edu

ABSTRACT
In this paper we present work on the task of Native Language Identification (NLI). We present an alternative corpus to the ICLE which has been used in most work up until now. We believe that our corpus, TOEFL11, is more suitable for the task of NLI and will allow researchers to better compare systems and results. We show that many of the features that have been commonly used in this task generalize to new and larger corpora. In addition, we examine possible ways of increasing current system performance (e.g., additional features and feature combination methods), and achieve overall state-of-the-art results (accuracy of 90.1%) on the ICLE corpus using an ensemble classifier that includes previously examined features and a novel feature (n-gram language models). We also show that training on a large corpus and testing on a smaller one works well, but not vice versa. Finally, we show that system performance varies across proficiency scores.

KEYWORDS: Native Language Identification, Text Classification, Corpora.

Proceedings of COLING 2012: Technical Papers, pages 2585–2602,
COLING 2012, Mumbai, December 2012.

1 Introduction

One growing NLP field is that of Native Language Identification (NLI), which is the task of automatically identifying a speaker's first language based solely on the speaker's writing in another language. NLI can be useful for a number of applications. Native language is often used as a feature in machine learning approaches to authorship profiling (Estival et al., 2007), which is frequently used in forensic linguistics. NLI can also be used in educational settings to provide more targeted feedback to language learners about their errors (Chang et al., 2008; Dahlmeier and Ng, 2011; Rozovskaya and Roth, 2011). It is well known that speakers of different languages make different kinds of errors when learning a language (Swan and Smith, 2001). For example, a French speaker learning English might write sentence (1), which contains a verb tense error. On the other hand, a Japanese speaker learning English might make the verb tense error shown in (2). A writing tutor system which can detect the native language of the learner will be able to tailor the feedback about the error and contrast it with common properties of the learner's language.

(1) She knows that she hasn't achieve it completely.

(2) They also said to have great curiosity.

There has been a great deal of work on NLI in recent years. The methods employed have ranged from some combination of lexical, part-of-speech and n-gram features (Koppel et al., 2005), to syntactic features (Wong and Dras, 2011) including Tree Substitution Grammars (TSGs) (Swanson and Charniak, 2012), to topic models (Wong et al., 2011). Despite these research efforts, it has been somewhat hard to compare different approaches for a number of reasons.

The first difficulty is with the evaluation data set. Evaluating an NLI system requires a corpus containing texts in a language other than the native language of the writer. Because of a scarcity of such corpora, most work[1] has used the ICLEv2[2] for training and evaluation since it contains several hundred essays written by college-level English language learners. However, this corpus is quite small for training and testing statistical systems which makes it difficult to tell whether the systems that are developed can scale well to larger data sets or to different domains. The usability of the corpus is further compromised by idiosyncrasies in the data such as topic bias (as shown by Brooke and Hirst, 2011) and the occurrence of characters which only appear in essays written by speakers of certain languages. As a result, it is hard to draw conclusions about which features actually perform best.

A second problem is that there is no consensus on the scope of the evaluation. The ICLE contains English essays written by native speakers of 16 languages. Typically a subset of 7 languages is used in the evaluations, although more recently some work has reported results for a larger set. Moreover, when researchers report results for 7 languages, they are not always reporting on the same 7 languages. For example, in the work of Wong and Dras (2011) the 7 native languages (L1s) are Bulgarian, Chinese, Czech, French, Japanese, Russian, and Spanish. Whereas in Brooke and Hirst (2012), Italian and Polish are used instead of Bulgarian and Czech. In addition, different researchers have split the corpus in different ways when training and evaluating their systems, making it even more difficult to compare results across experiments.

[1] Note that Kochmar (2011) used a subsection of the Cambridge Learner Corpus.

[2] Throughout this paper, we will refer to ICLE version 2 as ICLE.

In this paper, we first provide an automatic method for extracting data from the ICLE corpus to remove some of the corpus-specific idiosyncracies that automatic Native Language Identification classifiers currently learn from. We call this modified version of the ICLEv2, ICLE-NLI. Second, we introduce a new data set, TOEFL11 (Blanchard et al., to appear), which is roughly twice the size of the ICLE and has more essays per L1 than the ICLE. We argue for the use of TOEFL11, which will be made publicly available, as a common evaluation resource for this task. Next, we use these two new resources (ICLE-NLI and TOEFL11) to address the following research questions:

1. Are there methods previously unexplored in the literature that can be used to improve performance? (Section 5.1)
2. Do features commonly used in prior work generalize to different corpora? (Section 5.1)
3. What is the effect of training on one corpus and then testing on another? (Section 5.2)
4. What is the effect of larger training data on the performance of these features? (Section 5.3)
5. How widely do results vary across levels of writing proficiency? (Section 5.4)

In Section 2, we discuss previous approaches to NLI and also how there is a need for greater standardization of corpora used and evaluation practices. We present the corpora used in this study in Section 3, and our system and the features we investigate in Section 4. In Section 5 we discuss the results and the five research questions above. Our best system achieves state-of-the-art accuracy (90.1%) on the ICLE-NLI corpus, surpassing the previously reported best accuracy of 81.7% (Wong and Dras, 2011) on ICLE.

2 Related Work

The work of Koppel et al. (2005) set the stage (and probably a high bar as well in terms of performance) for much of the NLI research in the past few years. Their work investigated features from NLP and Second Language Acquisition including character and POS n-grams, content and function words, and spelling and grammatical errors. The features were evaluated on a subsection of the ICLE corpus consisting of essays sampled from 5 L1s (Russian, Czech, Bulgarian, French and Spanish) with 10-fold cross validation. The researchers found that by combining all of the features using a SVM, they could achieve an accuracy of 80.2%. Tsur and Rappoport (2007) continued this work by investigating why character n-grams alone performed so well (66%). As in the work of Wong and Dras (2011), we use an approximation of the Koppel et al. (2005) features both as a baseline system and as a base feature set to which we add our own features.

The notion that different learners tend to exhibit different grammatical error patterns was further explored by Kochmar (2011). Instead of the ICLE corpus, English learner essays from the Cambridge Learner Corpus[3] were used, specifically essays written by test-takers with Romance and Germanic native languages. In this work, a SVM was used to classify these essays on the basis of lexical and parse features as well as manually marked grammatical and spelling errors.

[3] http://www.cup.cam.ac.uk/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus

One of the main conclusions was that character and POS n-grams were the most powerful features. Since this work used a different evaluation corpus and different L1s, it is hard to compare to other studies.

Wong and Dras (2009) based their NLI system on the contrastive analysis hypothesis: that differences between a writer's L1 and the language they are trying to write in are caused by differences between the two languages. Often, characteristics of the writer's L1 are carried over into the target language. They investigated the impact of three common ESL error types: subject-verb agreement, noun-number agreement and determiner errors and used 7 languages from ICLE: the 5 used in the Koppel et al. (2005) study in addition to Chinese and Japanese. While the determiner error feature did seem informative, performance did not improve when it was combined with a baseline model of lexical features.

Recently, more complex syntactic features have been proposed to better model the structural differences in learner writing. Wong and Dras (2011) extended the Koppel et al. work by incorporating production rules from two parsers as well as reranking features into the classification framework. In line with their prior work, they address the 7-way NLI task (Bulgarian, Chinese, Czech, French, Japanese, Russian and Spanish) by selecting 95 essays for each L1 and doing 5-fold cross validation with 70 essays selected for training and 25 for testing. This particular evaluation framework has been adopted by others in the field, such as Swanson and Charniak (2012). The combination of parse production rules with the Koppel et al. lexical features achieved a performance of 81.71% on the 7-way NLI task, currently the highest reported in the literature for that set of 7 in the ICLE. Swanson and Charniak (2012) experiment with various Tree Substitution Grammars (TSGs) on the 7-way NLI task and achieve an accuracy of 78.4%. In our work, we also experiment with the use of the Post and Gildea (2009) TSG fragments for the NLI task. Bergsma et al. (2012) tackled the related problem of classifying text as either written by a native speaker or a non native speaker of English. They used TSGs in conjunction with other stylometric features to achieve an F-score of 91.6.

Wong et al. (2011) also investigated the use of Latent Dirichlet Analysis (LDA) to cluster features and then adaptor grammars (Wong et al., 2012) to better capture arbitrary length n-gram sequences over tags and words. While the LDA approach did not improve performance over a baseline model of lexical features, the adaptor grammar approach scored close to state-of-the-art, around 75% accuracy.

Finally, Brooke and Hirst (2012) explored the effect of adding more training data on NLI classifier performance. In one experiment, English sentences written by Chinese and Japanese native speakers were scraped from a language-exchange social networking website (lang8)[4] and used to augment a classifier which used different n-gram features and function words. The effect was that for those two languages, increasing the training data from 200 texts to 5,000 improved overall performance by 30% (from 59.8% to 89.8%). This work sets the tone for the resource contribution of this paper: the creation of a publicly available corpus of ESL essays.

3 Data

In this work, we evaluated our NLI systems on four corpora: a subset of the ICLE (Granger et al., 2009) and three samples from a collection of essays written by non native English speakers as part of a high stakes college-entrance test (TOEFL®). Each corpus is described in detail below.

[4] http://www.lang-8.com/

3.1 ICLE Corpus Description

Currently, the only widely available corpus of non native English that is annotated for native language is the International Corpus of Learner English (ICLE) (Granger et al., 2009). It is a large collection of 6,085 essays written by university undergraduates of advanced proficiency.[5] The intent of the project was to produce for corpus linguistics a relatively large corpus that "shared a large number of task variables, notably in terms of medium (writing), genre (academic essay), field (general English rather than English for Specific Purposes) and length (between 500 and 1,000 words)" (Granger et al., 2009). Because these were the only factors the designers controlled for, there is substantial variation of other factors such as the number of essays per L1 and the number of languages covering each topic (see Table 1). The current version of the corpus, ICLEv2, consists of 6,085 essays for 1,302 prompts, which we have manually clustered into 736 topics.

Language	**Unique Topics**	**Total Topics**	**Essays on Unique Topics**	**Total Essays**
Bulgarian (Bul.)	0	4	0	302
Chinese (Chi.)	16	27	542	982
Czech (Cze.)	29	49	46	243
French (Fre.)	6	21	33	347
Japanese (Jap.)	119	132	336	366
Russian (Rus.)	2	17	2	276
Spanish (Spa.)	14	32	60	251

Table 1: ICLE counts of topics and essays for seven L1s commonly used in the NLI task.

3.2 Using the ICLE for NLI

Because techniques for NLI have looked for patterns in the data not only at the lexical level but also at the character level, any unintended lexical or character patterns correlated with L1 are likely to be weighted heavily by a statistical classifier. Therefore, it is important that the corpus is free of such patterns.[6] After examining the ICLE in detail, we discovered that there are two classes of confounding patterns for the task of NLI:

1. A variety of character encoding errors and annotations (both erroneous and correct) occur predominantly in certain languages and not in others. Because of the techniques and features commonly used for NLI, these issues can impact system evaluation.

2. The topics the authors write about are not evenly distributed across languages. In fact, there are many topics for which all the authors are native speakers of a single L1 (see Table 1). This topic bias is also observed and discussed in Brooke and Hirst (2011). The bias is problematic for machine learning approaches to NLI because it could cause classifiers to conflate the tasks of topic identification and NLI. For example, only Chinese

[5] http://www.uclouvain.be/en-cecl-icle.html

[6] We do acknowledge that there are two ways of evaluating NLI systems. The first one, and the approach we are taking here, is that all characters are encoded in the same way. The second way of evaluating is to actually include any patterns that may arise from using a certain keyboard or encoding scheme that writers with a certain L1 tend to use. The first method essentially focuses on solely on linguistic patterns, while the second is more application-driven and focuses on both linguistic and extra-linguistic patterns.

authors responded to the prompt “Discuss the advantages and disadvantages of using credit cards,” and consequently all of those essays contain the $ character, whereas almost no other languages’ essays do.

To address the first set of problems we (a) removed the header information from all documents, (b) removed all instances of the codes used by the annotators to indicate deleted references, quotations, or illegible words, (c) converted all non-ASCII characters to their closest ASCII equivalent using the Unidecode Python module,[7] (d) fixed characters that resulted from encoding errors by replacing them with the appropriate character, (e) deleted long quotes that were surrounded by <</>> instead of the usual English quote character ", and (f) removed the duplicate essay from `RUMO7057.txt`

To remove as much topic bias as possible from our ICLE dataset, we used a specific sample of the corpus that did not contain any topics that were found in one and only one native language group. The only exception to this was the sample for Japanese, where most of the topics were unique to that language group. We chose to leave the unique Japanese prompts in the sample because the alternative would have been to remove all of the Japanese essays from consideration. For the evaluations with ICLE discussed later in this paper (Section 5), we use this modified version of the ICLE. We provide a script which will automatically modify ICLE to address the issues above. The transformed version of this corpus is called ICLE-NLI.

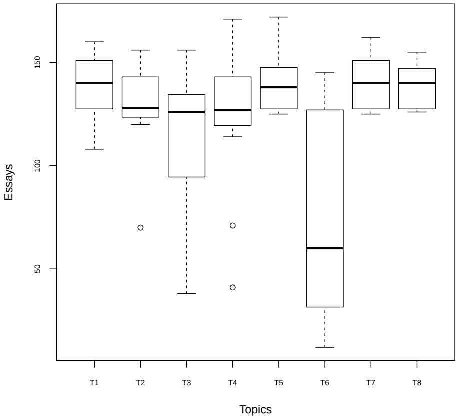

Figure 1: TOEFL11: Language distribution for each topic

3.3 TOEFL11: A New Corpus for NLI

While the modifications do address many of the issues with the ICLE corpus, it still has problems of small data size and some remaining topic bias. We constructed a new corpus of non native English writing called TOEFL11 which will be released through the LDC in 2013.

[7] http://code.zemanta.com/tsolc/unidecode/

This corpus consists of essays written by non native English speakers during a high stakes college-entrance test. It contains 1,000 essays per language sampled as evenly as possible from 8 topics along with author proficiency scores (low/medium/high) for each essay determined by assessment specialists. The distributions of the number of essays from each L1 group for each topic are shown in Figure 1. The 11 native languages covered by our corpus are: Arabic (Ara.), Chinese (Chi.), French (Fre.), German (Ger.), Hindi (Hin.), Italian (Ita.), Japanese (Jap.), Korean (Kor.), Spanish (Spa.), Telugu (Tel.), and Turkish (Tur.).

Corpus	**Languages (bold = common)**	**Essays per L1**	**Total Essays**	**Topics**	**Avg. Langs per Topic**	**Avg. Words per Essay**
ICLE-NLI	Bul., **Chi.**, Cze., **Fre.**, **Jap.**, Rus., and **Spa.**	110	770	76	1.4	666
TOEFL7	Bul., **Chi.**, Cze., **Fre.**, **Jap.**, Rus., and **Spa.**	659	4,613	79	6.5	252
TOEFL11	Ara., **Chi.**, **Fre.**, Ger., Hin., Ita., **Jap.**, Kor., **Spa.**, Tel., Tur.	1,000	11,000	8	11	315
TOEFL11-Big	Ara., **Chi.**, **Fre.**, Ger., Hin., Ita., **Jap.**, Kor., **Spa.**, Tel., Tur.	7,954	87,502	76	11	256

Table 2: Properties of NLI Corpora

We also compiled a larger dataset to investigate the effects of large amounts of training data on NLI accuracy. This corpus, which we will refer to as TOEFL11-Big, contains 87,502 essays from the same 11 languages as our public corpus, with between 7,900 and 7,983 essays per language. Again we sampled as evenly as possible across topics, but this time from a larger pool of 76 topics. There is no overlap in data between TOEFL11 and TOEFL11-Big.

To allow for a more direct comparison with previous ICLE results, we compiled a third corpus, henceforth TOEFL7, that uses the same 7 native languages that are most frequently used for NLI with the ICLE: Bulgarian, Chinese, Czech, French, Japanese, Russian, and Spanish. We used 659 essays per L1, sampled from a pool of 79 topics. In comparison, ICLE work typically uses 70 or 110 essays per L1.

Table 2 summarizes the four corpora to be used for training and testing in this paper. Of the three new corpora introduced in this paper, TOEFL11 will be made public.

4 System

Following previous work, we treat the problem of native language identification as a classification task. We train a native language identification system using a logistic regression model with L1-regularization.[8] We carry out 10-fold cross validation on all of the TOEFL® corpora and 5-fold cross-validation on the ICLE-NLI corpus (following Wong and Dras (2011)). In addition to building individual classifiers for each of our features, we also experiment with two ways of combining them. **Simple Combination** involves combining the features into one large set. In

[8] For all of our experiments we use the liblinear implementation (Fan et al., 2008) available from `http://www.csie.ntu.edu.tw/~cjlin/liblinear/` with the L1-regularized logistic regression solver and default parameters.

Character n-grams	All unigrams and bigrams present in each essay.
Function words	Function word counts based on the same list used by Koppel et al. (2005).
Part-Of-Speech (POS) bigrams	All POS bigrams present in each essay. Tags were obtained using the Stanford Tagger (Toutanova et al., 2003).
Spelling errors	Spelling errors returned by MS Word 2007 based on the list of error types used by Koppel et al. (2005). In the Koppel et al. (2005) work, spelling errors were derived using a pre-2003 version of MS Word.
Word n-grams	For unigrams, we restrict the list to correctly-spelled content words to prevent overlap with other features. For word bigrams, we do not filter the list.
Writing quality features	Counts and binary features quantifying different aspects of writing quality such as grammatical errors, style, discourse and vocabulary level. These features were derived using a proprietary automatic essay scoring engine (Attali and Burstein, 2006). Koppel et al. (2005) used a pre-2003 version of MS Word to find grammatical errors.
Tree Substitution Grammar Fragments	Counts and binary features corresponding to TSG fragments. These were extracted using the software described by Post and Gildea (2009) available from `https://github.com/mjpost/dptsg`.
Stanford Dependencies	Counts of basic dependencies extracted using the Stanford Parser (de Marneffe et al., 2006). Also included are variations of dependencies where lemmas were replaced by POS tags.
Language Model	Perplexity scores from 5-gram language models, one for each language in the corpus. Language Models were trained using the IRSTLM toolkit (Federico et al., 2008) .

Figure 2: A summary of all features used in our classifier

Ensemble Combination, which has not been explored for this task, each individual feature is trained as its own classification system and the predictions from that system, along with the scores (either probabilities or perplexity scores) are used in the final ensemble model to predict the native language.[9]

4.1 Features Used

The features used in our system are summarized in Figure 2. In order to be able to examine the performance of features across corpora, we implemented many of the features commonly used in the literature, as well as a novel language-model-based feature.

To begin with, we implement the features described by Koppel et al. (2005). The main classes of features in that paper were: character n-grams, function words, parts of speech, spelling errors, and writing quality features. We implement these features, roughly corresponding to the original work. One key difference is that we do not restrict our n-gram features to only include n-grams that occurred a certain number of times; we include all n-gram features as input to the classifier unless otherwise specified. For each feature, we implement both the binary and frequency variants. This set of binary and frequency features is combined to form the "Koppel Baseline."

We also combine Koppel's features with content words and word bigrams to produce a model

[9] Note that the splits used for all the individual classifier training and testing were identical.

that includes most of the simple features previous researchers have used for NLI. We only include counts of correctly spelled content words for the unigrams to avoid overlap with the other Koppel Baseline features.

Tree Substitution Grammar features have been used by previous researchers for this task (e.g., Swanson and Charniak, 2012). We take the fragments as generated by the Post and Gildea (2009) system and use those as fragments in our classification system. The Stanford dependency features are a variation on the syntactic features proposed previously (e.g., Kochmar, 2011; Wong and Dras, 2011; Post, 2011). We automatically extract all basic dependencies for each essay and consider each dependency to be a feature. This yields features of the form `nsubj(saw, dog)`, `dobj(saw, cat)`. We also carry out a backoff transformation based on part-of-speech, and for the two dependencies previously listed, would also consider the following features: `nsubj(VBD, dog)`, `dobj(VBD, cat)`, `nsubj(saw, NN)`, `dobj(saw, NN)`, `nsubj(VBD, NN)`, `dobj(VBD, NN)`.[10] We also consider the corresponding unlabeled dependencies as features. For almost all features we use both count-based (relative frequency for all but the writing quality features) and binary (presence/absence) features. We do not use binary features for the Stanford dependencies.

In addition to the previously described features, we also propose the use of language model perplexity scores. Surprisingly, language models, to our knowledge, have not been used for native language identification. We hypothesize that previous researchers may have avoided them because of the topic bias inherent in the ICLE or because there may not be enough data to build reliable models. Jarvis et al. (2012) showed that using higher-order *n*-gram features did not help for the task of L1 identification in the ICLE with 12 L1s and 10-fold cross-validation. We take a slightly different approach and train a 5-gram language model (with Witten-Bell smoothing) for *each* language in the corpus. We then apply each language model to each essay in the test data and choose as the prediction, the language model with the lowest perplexity.

5 Results

In this section, we discuss evaluations of different NLI systems under a variety of conditions to best answer the original research questions. In Section 5.1, we address the first two research questions by investigating the effectiveness of typical NLI features, as well as a novel feature (n-gram language models) across different corpora. In addition, we show that different feature combination methods can further improve system performance. In Section 5.2, we discuss the impact of training on one corpora and testing another. In Section 5.3, we discuss the impact of increasing the amount of training data on system performance. Finally, in Section 5.4 we show that the proficiency level of a writer should be taken into account when designing features and reporting results.

5.1 General Feature Discussion

This section reports the results of our experiments on the four corpora described in Section 3 to address the first two research questions posed. For each experiment we carried out cross validation, training on one portion of the data and testing on the remainder. We measure the performance of the classifier in terms of accuracy, i.e. the percentage of correctly classified languages in the corpus. Table 3 gives the results for all experiments. As a baseline, we take

[10] We ignore the index information provided by the Stanford Parser when storing the dependencies, however they are used when applying the part-of-speech backoff transformation.

our approximation of the Koppel et al. (2005) features. We also report results for features derived from Tree Substitution Grammar fragments, Stanford dependencies, and our novel language model feature. Finally the table shows that the ensemble method of combining features performs best, often significantly better than the simple combination of all features into one model.

On the ICLE-NLI corpus, our best ensemble model achieves an accuracy of 90.1% for a 7-way classification task. This is substantially higher than any previously reported results on the ICLE. Note that our experiments were carried out on a cleaned up version of the ICLE corpus (in an attempt to remove some of the biases, making the task more difficult). As noted in Section 2, it is very difficult to compare results for this task on the ICLE corpus because of the many different experimental procedures used by different researchers. However, despite these caveats, we believe that our system achieves state-of-the-art performance on this corpus for this task. Looking at the results for individual features, it is clear that the Koppel features alone do a very good job at predicting. Interestingly, the language model-based classifier also performs very well. The syntax-based features alone are also quite predictive, and our experiments show that the slightly more abstract representation of the syntactic structures within essays that can be captured by the Stanford dependencies is also a powerful predictor.

The results for TOEFL7, which contains the same languages as the ICLE-NLI corpus with a similar number of total topics but is around 6 times larger, show that the features and combinations that worked extremely well for the ICLE-NLI corpus, do not perform as well on this new corpus. The best result is achieved by our ensemble model with an accuracy of 70.9%. Again we see that the baseline Koppel features are very strong, although the language model-based classifier does not outperform the syntax-based models on this corpus. We believe that this indicates that while we may be reaching the upper bounds of performance on the ICLE-NLI corpus, there still remains plenty of room to improve current models and develop them to be able to perform well across corpora.

Interestingly, all of the combined models (and most of the individual ones) perform substantially worse on TOEFL7 than on any of the other corpora. There are two properties of the TOEFL7 corpus that are likely the cause for this performance disparity: (a) essay length and (b) topic distribution. The essays in the ICLE-NLI are on average 666 words long (see Table 2), whereas the TOEFL7 essays are an average of 252 words long. Therefore, while there are more total essays in the TOEFL7 corpus to use for training than in the ICLE-NLI sample, there are fewer potential occurrences of each useful feature per essay, which means the binary versions of the features are less helpful. This length disparity does not exist between the TOEFL7 and TOEFL11 data sets, but the distribution of topics is very different, which may contribute to the difference in performance. The topic distributions for TOEFL11 and TOEFL11-Big were made as even as possible, but we did not have access to enough of some of the ICLE-NLI languages to be able to sample TOEFL7 as evenly.

Our performance on the TOEFL11 corpus, developed specifically for the task of native language identification, are encouraging. Our best ensemble model achieves 80.9% accuracy, and the performance of the individual features follows a pattern similar to the ICLE-NLI and TOEFL7 evaluations. These results show that the features and their combinations do scale to larger corpora (1000 essays per language, 11 native languages)

Finally, on our largest corpus, TOEFL11-Big we see the same patterns as on the other corpora. Here, the increased data size does seem to lead to some small improvements in performance

Features	**ICLE-NLI**	**TOEFL7**	**TOEFL11**	**TOEFL11-Big**
	(7-way)	**(7-way)**	**(11-way)**	**(11-way)**
Random baseline	14.3	14.3	9.1	9.1
Koppel baseline	80.0	58.0	65.9	67.3
character unigrams	20.1	23.2	23.8	28.1
character unigrams (binary)	46.1	29.2	26.6	24.3
character bigrams	14.5	26.0	25.0	39.2
character bigrams (binary)	70.3	44.4	52.2	51.9
function words	23.1	25.5	27.2	42.3
function words (binary)	57.7	38.8	45.9	43.4
POS bigrams	25.6	31.3	26.8	41.4
POS bigrams (binary)	54.8	42.0	38.5	46.0
spelling	29.9	31.8	31.1	30.7
spelling (binary)	29.5	31.3	29.6	30.1
writing quality	57.4	35.6	37.2	32.8
writing quality (binary)	40.6	27.8	26.0	24.3
Koppel + word n-grams	82.9	67.5	76.6	81.4
word unigrams	22.7	18.6	21.3	44.9
word unigrams (binary)	76.5	52.7	64.9	63.5
word bigrams	14.3	19.0	22.2	41.7
word bigrams (binary)	72.7	53.9	67.9	77.1
5-gram language models	80.8	53.4	73.9	74.4
tree subst. grammar frags.	74.4	55.6	62.6	64.3
stanford dependencies	77.1	59.3	70.9	76.7
simple combination	82.6	70.5	76.0	80.9
ensemble	**90.1**	**70.9**	**80.9**	**84.6**

Table 3: Cross-validation results for all systems on each corpus; accuracy in %

for some features, but not all. Particularly, some of the binary features perform worse on the TOEFL11-Big corpus compared to the smaller TOEFL11 corpus. The syntax-based features do seem to be one of the stronger features given the extra data.

We notice that the performance of the language-model feature is inconsistent across corpora. It remains unclear why this is the case, but we hypothesize that it is partly due to topic distributions. We will continue to explore the cause of this inconsistency. Noteworthy, is the fact that word unigrams/bigrams are generally one of the strongest baselines. Our language-model feature is a more complex version of these features. These simple features alone perform about as well as the entire set of features that goes into the Koppel baseline. One reason previous researchers had given for avoiding these word-level features was because it was felt that they were unfairly advantaged because of the topic bias in ICLE. Our experiments show, however, that even in corpora where there is no topic bias (TOEFL11), these word-level features remain predictive.

5.2 Cross-Corpus Evaluation

The experiments in Section 5.1 all involved cross-validation on one corpus. A remaining issue to address is the question of whether a system trained on one corpus can generalize to another. Brooke and Hirst (2011) reject cross-validation as a means of evaluating NLI systems on corpora that are heavily topic biased and instead train their system on a large amount of web-scraped data.

In our study of the effect of training on one corpus and testing on another, we carry out experiments on pairs of corpora that consist of the same sets of languages. We evaluate first on the ICLE-NLI vs TOEFL7 corpora (7 languages) and second on the TOEFL11 vs TOEFL11-Big corpora (11 languages). The main argument for carrying out this evaluation as proposed by Brooke and Hirst (2011) is to circumvent the issue of topic bias. We believe that these pairs of corpora are composed of sufficiently different topics that there should not be significant overlap.

For the ICLE-NLI vs TOEFL7 experiment, we train a classifier using the combined set of features listed in Figure 2 apart from the language model features. The results are reported in Table 4. The results generally show lower performance than the cross-validation experiments. In particular, the system trained on the ICLE-NLI data set does not generalize at all to the TOEFL7 data set. Training on TOEFL-Big and testing on TOEFL11 shows that the large amount of training data available in the TOEFL11-Big corpus leads to similar performance as the cross-validation result on TOEFL11. In general it seems that training on a larger corpus and testing on a smaller one works reasonably well, however training on a small corpus and testing on a larger one does not yield good results with our feature set. It remains for future work to determine which individual features can generalize well in this scenario.

Train	Test	Accuracy(%)
ICLE-NLI	TOEFL7	26.6
TOEFL7	ICLE-NLI	67.4
TOEFL11	TOEFL11-Big	35.4
TOEFL11-Big	TOEFL11	79.2

Table 4: Results for Cross-corpus evaluation

5.3 Corpus size

The relatively small size of the portion of the ICLE corpus traditionally used for the task of NLI has been a major criticism (Brooke and Hirst, 2011). Since our TOEFL11-Big corpus is several orders of magnitude larger than this, we are in a favorable position to be able to examine the impact of adding additional training data to NLI classifiers. In Brooke and Hirst (2011), they automatically gleaned additional training data by scraping online blogs. Our large data set is composed of essays written by learners of English and the native language self-reported by the learners.

Figure 3 shows the learning curve for an ensemble classifier on our TOEFL11-Big data set. When increasing the size of the data set, we add the same number of essays for each language, where possible. The graph shows a steep rise at the beginning, but as more data is added, the increase in performance is low and appears to be leveling off. The graph shows that the 11,000 essays in our publicly available data set should be a large enough number to be able to train classifiers with high accuracy, but that there could also be performance improvements with a somewhat larger dataset.

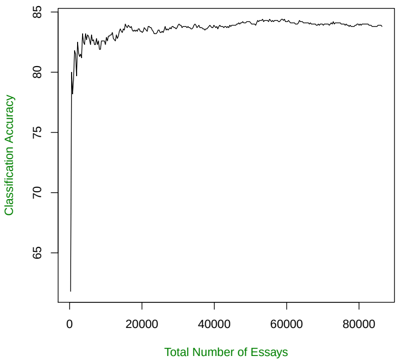

Figure 3: 11-way Classifier accuracy when trained on increasing amounts of data

5.4 Proficiency-based Evaluation

One issue that has not been thoroughly addressed in the field is the effect of writer proficiency on system performance. If a corpus is composed of highly proficient writers, then one would expect the effects of features such as spelling and grammatical errors to be less useful since they would be less common. Conversely, if the writer is of moderate or low proficiency, error features may be a larger contributor to classifier performance while more complex features may be less so due to data sparsity. So, when designing an NLI system and evaluating it, it is important to take into account the proficiency of the writer. In a pilot study, Jarvis and Paquot (2012) found that it is necessary to control for proficiency since it impacts the final discriminative analysis. They used trained assessment specialists to manually rate 223 ICLE essays written by learners of 3 L1s: Spanish, German and French. They found that the Spanish essays tended to have lower proficiency scores than the other two and were thus more easily distinguishable using the standard error techniques.

To investigate the influence of author proficiency on classification accuracy, we examined the performance of our best ensemble model at the three proficiency scores. As can be seen in Table 5, accuracy is highest for essays with a medium proficiency score. One possible reason for this is that medium proficiency essays feature a significant number of errors that many of the features (e.g. spelling errors, POS bigrams, and writing quality) can be used for prediction, but not so many that the sentences are seriously ungrammatical and difficult to interpret, as is the case with low-scoring essays. It is not surprising that high-score essays are difficult for NLI because their near-native writing would contain substantially fewer predictive errors. An alternative explanation for the results in Table 5 is that they simply reflect the score distribution in the corpus. We will continue to examine these hypotheses in future work.

We believe that proficiency reporting should be a necessary "best practice" in the NLI field. As more learner corpora become available, such as TOEFL11, Cambridge Learner Corpus and lang8, including a breakdown of classifier performance by proficiency score will make it easier to compare and discuss results across corpora.

Proficiency Score	Accuracy (%)	Number of Essays
Low	82.8	1,201
Medium	86.1	5,964
High	79.8	3,835

Table 5: TOEFL11 Accuracy for Best Model by Proficiency Score

Conclusion

To date, it has been hard to interpret results or compare systems because of the reliance on the ICLE corpus. In this paper, we addressed these issues by providing two resources: a modified version of the ICLEv2 (ICLE-NLI) and a larger corpus of essays that is more balanced across topics (TOEFL11). It is our hope that making these resources publicly available will help foster better standardization in the growing field of NLI.

Using these two corpora, we were able to draw the following conclusions:

1. Many of the trends found in previous work on the ICLE do generalize to other corpora, such as the power of the Koppel et al. (2005) features and word n-gram frequencies.
2. N-gram language models, which had been heretofore unexplored in this work, perform comparatively well for all corpora we examine.
3. Training on a large corpus and testing on a smaller corpus works well, but classifiers trained on smaller corpora do not appear to generalize well.
4. Combining multiple features in an ensemble classifier yields significantly greater classification accuracy than simply using including all of the features in one large classifier for all corpora we examine.
5. For the ICLE, the ensemble method has an accuracy of 90.1%, which is higher than the previously reported best accuracy of 81.7%.
6. Classification accuracy varies across proficiency levels, however further research is required in order to be able to explain the reasons for this.

Acknowledgments

The authors wish to thank Ben Swanson, Julian Brooke, Matt Post, Jojo Wong and Sylviane Granger for stimulating discussions related to this work. We would also like to thank the anonymous COLING reviewers, Jill Burstein and Nitin Madnani for their helpful comments and suggestions for improving earlier versions of this paper. We are grateful to Derrick Higgins, Anthony Ostrander, Xiaoming Xi and Robbie Kantor for their help in making the public release of the TOEFL11 dataset possible.

References

Attali, Y. and Burstein, J. (2006). Automated Essay Scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment*, 4(3). Available from http://www.jtla.org.

Bergsma, S., Post, M., and Yarowsky, D. (2012). Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337, Montréal, Canada. Association for Computational Linguistics.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. TOEFL11: A Corpus of Non-Native English. Research report (to appear), Educational Testing Service.

Brooke, J. and Hirst, G. (2011). Native language detection with 'cheap' learner corpora. In *Proceedings, Conference on Learner Corpus Research*, Louvain-la-Neuve. Presses universitaires de Louvain.

Brooke, J. and Hirst, G. (2012). Measuring interlanguage: Native language identification with l1-influence metrics. In *Proceedings, 8th ELRA Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul.

Chang, Y.-C., Chang, J. S., Chen, H.-J., and Liou, H.-C. (2008). An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpusbased NLP technology. *Computer Assisted Language Learning*, 21(3):283–299.

Dahlmeier, D. and Ng, H. T. (2011). Correcting Semantic Collocation Errors with L1-induced Paraphrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Edinburgh, Scotland, UK. Association for Computational Linguistics.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 449–454, Genoa, Italy.

Estival, D., Gaustad, T., Pham, S.-B., Radford, W., and Hutchinson, B. (2007). Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 263–272.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Federico, M., Bertoldi, N., and Cettolo, M. (2008). IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proceedings of Interspeech*, pages 1618–1621, Brisbane, Australia.

Granger, S., Dagneaux, E., and Meunier, F. (2009). *The International Corpus of Learner English: Handbook and CD-ROM, version 2*. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

Jarvis, S., Castañeda-Jiménez, G., and Nielsen, R. (2012). *Approaching Language Transfer through Text Classification*, chapter Detecting L2 Writers' L1s on the Basis of Their Lexical Styles, pages 34–70. Multilingual Matters.

Jarvis, S. and Paquot, M. (2012). *Approaching Language Transfer through Text Classification*, chapter Error Patterns and Automatic L1 Identification, pages 127–153. Multilingual Matters.

Kochmar, E. (2011). Identification of a writer's native language by error analysis. Master's thesis, University of Cambridge.

Koppel, M., Schler, J., and Zigdon, K. (2005). Automatically determining an anonymous author's native language. In *ISI*, pages 209–217.

Post, M. (2011). Judging grammaticality with tree substitution grammar derivations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 217–222, Portland, Oregon, USA. Association for Computational Linguistics.

Post, M. and Gildea, D. (2009). Bayesian Learning of a Tree Substitution Grammar. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 45–48, Suntec, Singapore. Association for Computational Linguistics.

Rozovskaya, A. and Roth, D. (2011). Algorithm Selection and Model Adaptation for ESL Correction Tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 924–933, Portland, Oregon, USA. Association for Computational Linguistics.

Swan, M. and Smith, B., editors (2001). *Learner English: A teacher's guide to interference and other problems*. Cambridge University Press, 2 edition.

Swanson, B. and Charniak, E. (2012). Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 193–197, Jeju Island, Korea. Association for Computational Linguistics.

Tetreault, J. R. and Chodorow, M. (2008). The Ups and Downs of Preposition Error Detection in ESL Writing. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 865–872, Manchester, UK. Coling 2008 Organizing Committee.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pages 252–259, Edmonton, Canada.

Tsur, O. and Rappoport, A. (2007). Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16, Prague, Czech Republic. Association for Computational Linguistics.

Wong, S.-M. J. and Dras, M. (2009). Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia.

Wong, S.-M. J. and Dras, M. (2011). Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Wong, S.-M. J., Dras, M., and Johnson, M. (2011). Topic modeling for native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 115–124, Canberra, Australia.

Wong, S.-M. J., Dras, M., and Johnson, M. (2012). Exploring adaptor grammars for native language identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 699–709, Jeju Island, Korea. Association for Computational Linguistics.